Supplementary Information: "Decomposing feature-level variation with Covariate Gaussian Process Latent Variable Models"

A ADD+INT kernel decomposition with mean-zero functional constraints

Suppose $f \sim \mathcal{GP}(0, k(\cdot))$ where f is a one-dimensional function and $k()$ is the squared exponential kernel,

$$k(x, y) = \sigma^2 \exp\left(-\frac{1}{2}\frac{(x-y)^2}{l^2}\right)$$

Following Durrande et al. [2012, 2013] we can construct a GP prior for f conditional on $\int_a^b f(t)dt = 0$.

Writing down the joint distribution of $f(x)$ and its integral $\int_a^b f(t)dt$ over some interval $[a, b]$,

$$\begin{pmatrix} f(x) \\ \int_a^b f(t)dt \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} k(x,x) & \int_a^b k(x,t)dt \\ \int_a^b k(t,x)dt & \int_a^b \int_a^b k(t,s)dtds \end{pmatrix}\right]$$

we can express the conditional distribution of $f(x)$ conditional on $\int_a^b f(t)dt = 0$, i.e. conditional on f being mean-zero. As a result,

$$f \mid (\int_a^b f(t)dt = 0) \sim \mathcal{GP}(0, \tilde{k}(\cdot)),$$

where

$$\tilde{k}(x, y) := k(x, y) - \frac{\int_a^b k(x,t)dt \int_a^b k(t,y)dt}{\int_a^b \int_a^b k(t,s)dtds}.$$

When $k(x, x')$ is the squared exponential kernel, these integrals have analytic forms,

$$\int_a^b k(x,t)dt = 0.5\sqrt{2\pi} l \sigma^2 \left(\text{erf}\left(\frac{b-x}{\sqrt{2}l}\right) - \text{erf}\left(\frac{a-x}{\sqrt{2}l}\right)\right)$$

$$\int_a^b \int_a^b k(t,s)dtds = \sqrt{2\pi} l \sigma^2 \left((a-b)\text{erf}\left(\frac{a-b}{\sqrt{2}l}\right) + \frac{\sqrt{2}}{\sqrt{\pi}} l(\exp(-\frac{(b-a)^2}{2l^2}) - 1)\right)$$

Now, being able to evaluate $\tilde{k}$, we can construct the mean-zero decomposition. We will formulate the ADD+INT kernel decomposition on the joint (z, x) space as follows

$$k((z,x),(z',x')) := \sigma_b^2 + \sigma_z^2 \tilde{k}_0(z,z') + \sigma_x^2 \tilde{k}_0(x,x') + \sigma_{zx}^2 \tilde{k}_0(z,z')\tilde{k}_0(x,x')$$

where $\tilde{k}_0$ is the mean-zero kernel as above with kernel variance set to 1.

References

Nicolas Durrande, David Ginsbourger, and Olivier Roustant. Additive covariance kernels for high-dimensional gaussian process modeling. *Annales de la Faculté des sciences de Toulouse : Mathématiques*, Ser. 6, 21(3):481–499, 2012. doi: 10.5802/afst.1342.

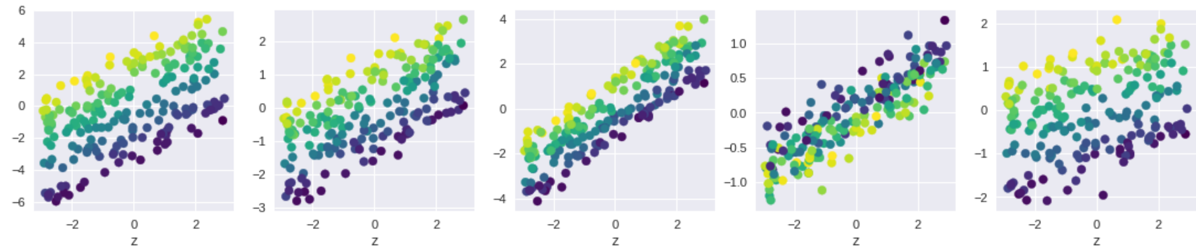
Nicolas Durrande, David Ginsbourger, Olivier Roustant, and Laurent Carraro. Anova kernels and rkhs of zero mean functions for model-based sensitivity analysis. *Journal of Multivariate Analysis*, 115:57–67, 2013.

B Additional experiments: Synthetic gene expression data

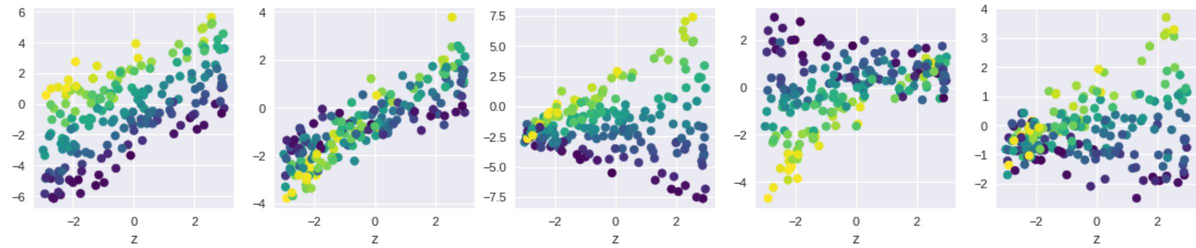
The "rings" and "pinwheel" exhibited highly non-linear patterns, whereas now we explore the behaviour of these models on more realistic synthetic examples, considering dependency structures that we expect to see in real gene expression data. For this purpose, we constructed synthetic data sets where features would exhibit (i) linear additive signal, (ii) linear interactions combined with additive signal, (iii) monotone dependency, and (iv) the latter combined with transient signals.

For the synthetic gene expression data, we generated the following 5-dimensional data sets ($\mathbf{z}$ on x-axis, $\mathbf{y}^{(j)}$ on y-axis, coloured by covariate $\mathbf{x}$).

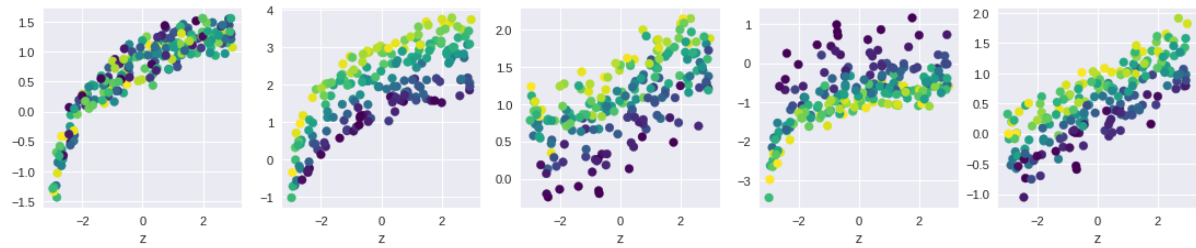
Linear ADD:



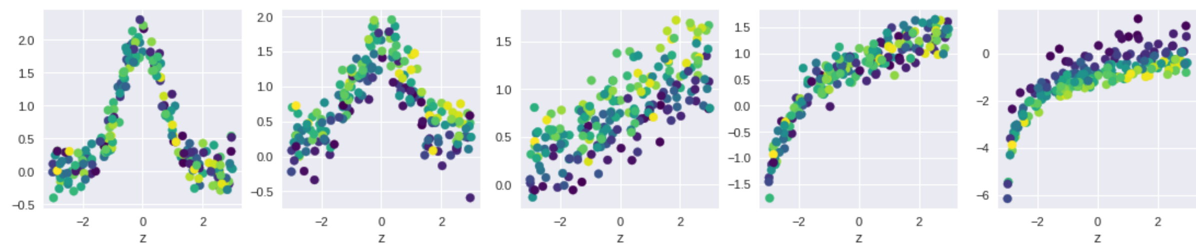
Linear ADD+INT:



Monotone:



Monotone + transient:



In these four scenarios, we measured how accurately do different models uncover the true underlying $\mathbf{z}$, results are summarised in Table 1.

Table 1: Under four data generation schemes (in columns), we compared the four models (in rows) in how accurately they recovered the true $\mathbf{z}$, displaying correlation between the true $\mathbf{z}$ and the inferred values.

	linear (ADD)	linear (INT)	monotone	monotone+ transient
LVR	**0.99**	**0.97**	0.91	0.84
GPLVM	0.81	0.82	0.80	0.91
sup-GPLVM	0.91	0.83	0.87	0.90
c-GPLVM	**0.99**	**0.97**	**0.96**	**0.97**

In the linear data simulations, the linear assumptions of latent variable regression (LVR) allows it to achieve near perfect recovering of the latent dimension but, despite the increased flexibility, so does c-GPLVM. When the data-generating mechanism is non-linear and including transient effects, the non-linear assumptions of GPLVM and sup-GPLVM exhibit their superiority over LVR but cannot recover the true latent structure as accurately as c-GPLVM.

C Details about the synthetic survival experiment (for section 5.2)

For the synthetic survival experiment, we generated a four-dimensional data set as shown below, where two features depend on $\mathbf{x}$ ($\mathbf{y}_1$ exhibits an interaction effect and $\mathbf{y}_2$ an additive effect), whereas the rest do not. We picked two data points with

- true $\mathbf{z} = -2.0$ and $\mathbf{x} = 1.5$
- true $\mathbf{z} = 1.0$ and $\mathbf{x} = 1.5$

to carry out the synthetic experiment.

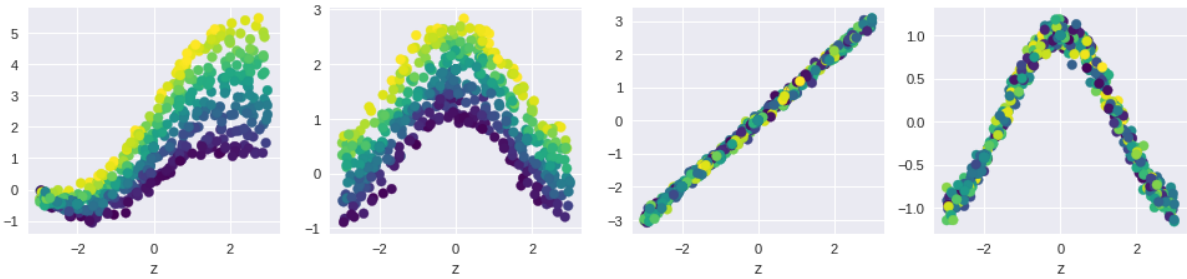


Figure 1: Data generation scheme (4-dimensional $\mathbf{y}$) for the synthetic censored-covariates experiment.

D Details about the survival experiment on TCGA data (for section 5.3)

As the posterior predictive from the Weibull regression does not take into account the fact that we have available censoring information, we have provided both the default predictions (which are not informed by censoring times) as well as the *conditional* predictions in Fig 2. We also quantify the prediction accuracy by repeatedly sampling from the inferred posteriors and show the distribution of mean squared error (MSE) in Figure 3.

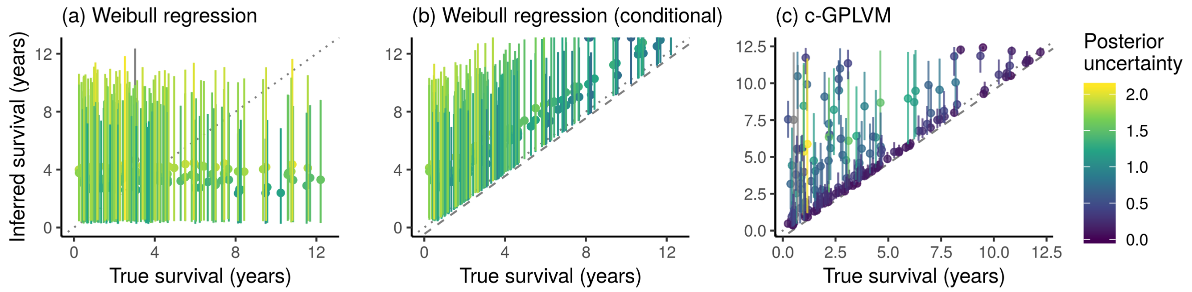


Figure 2

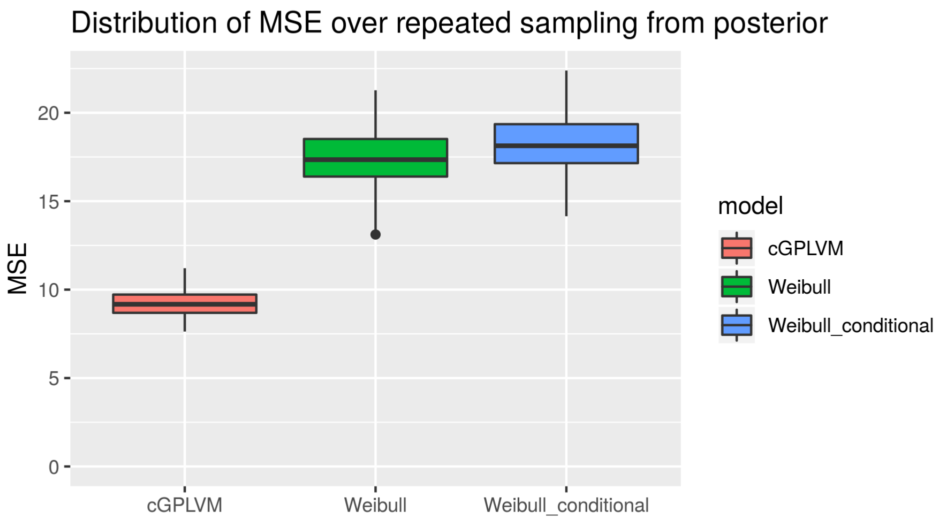


Figure 3: Mean squared prediction error for the TCGA artificial censoring experiment